BIMM-143
Introduction to Bioinformatics
http://thegrantlab.org/bimm143/

Bioinformatics databases

There are thousands of online bioinformatics databases. Here we list only a handful of the most commonly used (*i.e.* most highly cited in the literature) in the areas of biomolecular sequence, biomolecular structure, protein function and domain annotation, genome databases and model organisms.

Additional databases of potential interest can often be found by looking through the 'Nucleic Acid Research (NAR) Database Issue' available online. When considering a particular database remember that desirable features will likely include:

- Contains the data you are interested in.
- Allows fast data access.
- Provides annotation and curation of entries.
- Provides links to additional information (possibly in other databases).
- Allows you to make discoveries!

NCBI and EBI: Key database providers

The National Center for Biotechnology Information (**NCBI**) and European Bioinformatics Institute (**EBI**) are the most prominent online bioinformatics resource providers (both tools and databases).

Notable NCBI databases include:

- GenBank - an annotated collection of all publicly available DNA sequences.
- RefSeq - annotated set of non-redundant reference sequences (best representation of a sequence in their judgment) including genomic, transcript, and protein.
- PubMed - database of published biomedical literature (mostly abstracts).
- dbSNP - database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations of nucleotide sequences.

Notable EBI databases include

- ENA - a comprehensive record of DNA sequences. Contains the same sequences as GenBank (above) but offers different views and ways to navigate through the data.
- UniProt - the premier protein sequence database.
- Ensembl - genome databases for vertebrates and other eukaryotic species.
- PDBsum - pictorial database of 3D biomolecular structures in the Protein Data Bank (or PDB).

Biomolecular sequence databases

- GenBank - NCBI's nucleotide sequence database. Part of the 'International Nucleotide Database Collaboration' together with the ENA ('European Nucleotide Archive', from the EBI) and DDBJ (in Japan).
- RefSeq - The Reference Sequence collection constructed by NCBI to provide a comprehensive, integrated, non-redundant set of DNA, RNA sequences and protein products. It provides a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies and comparative analyses.
- UniGene - An Organized View of the Transcriptome created by NCBI. Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus, together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location.
- dbSNP - The database of single nucleotide polymorphism maintained by NCBI.
- UniProt - The main protein sequence database consisting of the protein 'KnowledgeBase' (UniProtKB), the sequence clusters (UniRef) and the sequence archive (UniParc).

Biomolecular structure databases

- PDB - The main repository of biomolecular structures maintained by the Research Collaboration for Structural Bioinformatics (RCSB). The same structures are also contained in PDBe, an EBI maintained version of the protein data bank that offers differing levels of annotation.
- SCOP - The database of Structure Classification of Proteins developed and maintained by Cambridge University.
- CATH - The database of protein structure 'Class, Architecture, Topology and Homologous superfamily' developed and maintained by University College, London.

Protein function and domain databases

- PFam - A database of protein families represented by multiple sequence alignments and hidden Markov models, constructed and maintained by the Sanger Institute, UK.
- Prosite - A database of protein domains, families and functional sites, created and maintained by the Swiss Institute of Bioinformatics.
- PRINTS - A database of protein fingerprints consisting of conserved motifs within a protein family, created and maintained by Manchester University, UK.
- BLOCKS - A database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins, created and maintained by the Fred Hutchinson Cancer Research Center, US.

- ProDom - A database of protein domain families automatically generated from UniProt sequence database, developed and maintained by the University Claude Bernard, France.
- HPA - A web site for the the human protein atlas which shows expression and localization of proteins in a large variety of normal human tissues, cancer cells and cell lines with the aid of immunohistochemistry images, developed and maintained by Proteome Resource Center, Sweden.

Genome databases and genome browsers

- ENSEMBL - The web server of the European eukaryotic genome resource developed by the EBI and Sanger Institute.
- UCSC Genome Information - The genome browser website containing the reference sequence and working draft assemblies for a large collection of genomes at the University of California at Santa Cruz (UCSC).
- NCBI Map Viewer - The NCBI genomic map viewer for the visualization of completed and ongoing genome sequence.
- NCBI Genome - The entry portal to various NCBI genomic biology tools and resources, including the Map Viewer, the Genome Project Database and the Plant Genomes Central, etc.
- NCBI Genome Information - The NCBI genomic information table lists the general information of genomes for all species.
- VISTA - A comprehensive suite of programs and databases for comparative analysis of genomic sequences.
- GOLD - Genomes Online Database, a comprehensive information resource for complete and ongoing genome sequencing projects with flowcharts and tables of statistical data.

Plant genome databases

- Phytozome - A tool for green plant comparative genomics, maintained by the Center for Integrative Genomics, Joint Genome Institute.
- Gramene - A curated open-source data resource for comparative genome analysis in the grasses including rice, maize, wheat, barley, sorghum etc, as well as other plants including arabidopsis, poplar and grape. Cross-species homology relationships can be found using information derived from genomic and EST sequencing, protein structure and function analysis, genetic and physical mapping, interpretation of biochemical pathways, gene and QTL localization and descriptions of phenotypic characters and mutations.
- TAIR - The Arabidopsis information resource maintained by Stanford University. It includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.

- AtENSEMBL - A genome browser for the commonly studied plant model organism Arabidopsis thaliana.
- Oryzabase - A comprehensive rice science database maintained by National Institute of Genetics, Japan. It contains genetic resource stock information, gene dictionary, chromosome maps, mutant images and fundamental knowledge of rice science.
- MaizeDB - The community database for biological information about the crop plant *Zea mays* ssp. *mays*, with genetic, genomic, sequence, gene product, functional characterization, literature reference.
- SoyBase - Integrating Genetics and Molecular Biology for Soybean Researchers.
- SGN - A collection of data resource of the Solanaceae species including tomoto, potato, peppper, eggplant, petunia, nicotiana.
- ICuGI - The web portal for the International Cucurbit Genomics Initiative including melon, cucumber, watermelon, pumpkin, etc.

Other genome databases

- PATRIC - the Bacterial Bioinformatics Resource Center, an information system designed to support the biomedical research community's work on bacterial infectious diseases via integration of vital pathogen information with rich data and analysis tools.
- GenoList - The bacterial genome database maintained at the Pasteur Institute.
- CyanoBase - The genome database for cyanobacteria developed by Kazusa Institute, Japan.
- Viral Genomes - the main page of NCBI viral genome information resource.
- GISAID - Global Initiative on Sharing Avian Influenza Data.
- OpenFlu - A database for human and animal influenza virus.
- NCBI Flu - NCBI Influenza Virus Resource with influenza genomic data and analysis tools.
- Plant Viruses - This site provides a central source of information about viruses, viroids and satellites of plants, fungi and protozoa.

Model organism focused database

- MGI - The international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease.
- ZFIN - The Zebrafish International Resource Center.
- Flybase - A comprehensive database of drosophila genes and genomes maintained by Indiana University.
- WormBase - The biology and genome resource of the *Caenorhabditis elegans* genome.

- SGD - The *Saccharomyces* Genome database.
- RGD - The Rat Genome Database at the Wisconsin University, to collect, consolidate, and integrate data generated from ongoing rat genetic and genomic research.
- XenBase - The African clawed frog *Xenopus laevis* and *Xenopus tropicalis* biology and genomics resource.

Cancer specific databases

- ICGC Data Portal - Tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.
- TGCA portal - Search, download, and analyze data sets generated by the 'Cancer Genome Atlas' (TCGA). It contains clinical information, genomic characterization data, and high level sequence analysis of tumor genomes.
- UCSC Cancer Genomics Browser - Interactively explore cancer genomics data and its associated clinical information.